A USER-CENTERED COMPARISON OF WEB SEARCH TOOLS

Pavel Braslavski, Anton Shishkin

Institute of Engineering Science, Yekaterinburg

{pb, whoarym}@imach.uran.ru

This study explores a user-centered approach to the comparative evaluation of the Web search tool ProThes against popular all-purpose search engines Yandex and Google. An original research design was developed. Data were collected from 12 volunteers who performed 48 search tasks in total. Main outcomes include: (1) search strategy supported through ProThes can be quite effective for focused Web search and (2) ProThes' interface and system performance must be improved.

Introduction

The growth of the Web leads to high popularity of the online search services. Web search becomes an important everyday activity of many professionals and casual Web-surfers. Meeting the demand, Web search engines (SEs) show superior productivity and extensive content coverage. Commercial SEs aim for satisfying as many Web surfers as possible and therefore employ modest user interfaces in addition to simple query syntax by default. Nevertheless users vary greatly in search expertise, command of languages, cultural background, etc. At the same time, query formulation, i.e. the transformation of a user's information need into a list of keywords, appears challenging for many searchers and remains an informal process to a great extent [1].

These problems invoke investigations into the users' behavior, actual information needs, search strategies, query formulation and re-formulation steps, etc. There are two distinct research approaches: query log analysis [3, 4, 5] and user-centered studies [7]. Each of the approaches has its advantages and shortcomings.

Earlier [2] we introduced ProThes, a tool for focused Web search that combines meta-search features, thesaurus-based query techniques, and graphical user interface (GUI) for query specification and thesaurus visualization (see project page for details: http://imach.uran.ru/prothes). ProThes currently communicates with Google (www.google.com) and Yandex (www.yandex.ru). A Russian-English thesaurus sample of the domain "Automated Optical Inspection of the Printed Circuit Boards (PCB)" was build manually from scratch. It consists mainly of PCB and computer vision related concepts. The thesaurus contains approximately 200 concept entries and 800 bilingual term entries. ProThes' GUI differs notably from standard SE interfaces (Figure 1).

After a pilot version of ProThes was developed, we needed to find out, whether the innovations have positive impact on search quality, comfort, and effectiveness. Since ProThes is a domain-specific search tool, moreover, a meta-searcher (i.e. it has no own search index) we cannot use standard test collections (like ROMIP/RIRES [6]). Similarly, we cannot conduct a comprehensive log analysis since only sparse and disembodied data are available. So we opted for user-centered approach. Our research design is similar to the one described in [7], expanded by comparative features analogously to experiments on evaluation of query-biased summarization [8, 9].

The goal of our research is twofold. First, we aim at developing a user-centered evaluation procedure for various search tools. Second, we want to evaluate different features of ProThes in order to define directions for future project development. Comparative data on user interaction with commercial SEs can be seen as a secondary outcome of the study.

This paper describes a research design, participants' characteristics and the collected data, followed by discussion of results and conclusion.

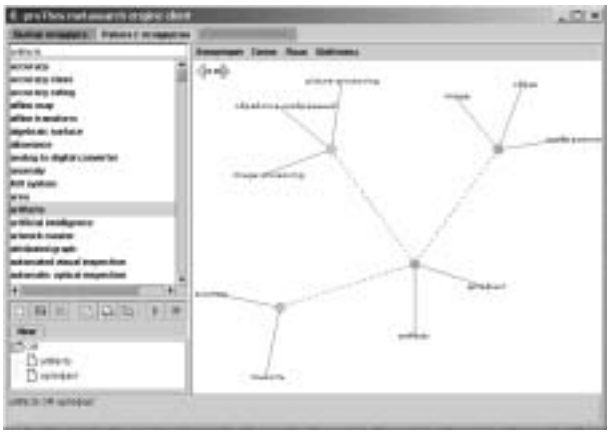

Fig. 1. ProThes' graphical user interface

Research design

The main parameters to be investigated within the study are usability of search tool and characteristics of user interaction with search tool. We use two methods for collecting data: (1) user pre- and post-search questionnaires and (2) search transactions logs. The research scheme is designed based on available time and equipment for data collection and processing.

The study consists of 12 experiments. In each experiment a user completes four search tasks (two using ProThes, and two using either Google or Yandex). Task pool consists of six tasks in the field of automatic opti-

cal inspection of printed circuit boards (see Appendix). There are three fact-finding (odd numbers) and three exploratory (open-ended) tasks (even numbers). Table 1 defines the combination of tasks, search tools, and execution order (we use P for ProThes, Y – for Yandex, and G – for Google). Thus, each task is performed eight times in total using different search tools (4 times – ProThes, 2 times – Google, and 2 times – Yandex).

Table 1. User/task/SE mapping

User	Task No.					
	1	**2**	**3**	**4**	**5**	**6**
1	G	G	P	P		
2		G	G	P	P	
3			G	G	P	P
4	P			G	G	P
5	P	P			G	G
6	G	P	P			G
7			P	P	Y	Y
8	Y			P	P	Y
9	Y	Y			P	P
10	P	Y	Y			P
11	P	P	Y	Y		
12		P	P	Y	Y	

The steps performed within each experimental routine are as follows:

1) Preparatory stage
   a) The instructor makes a brief introduction to study goals and the very experimental process.
   b) The user fills out a short questionnaire regarding familiarity with the English language and his/her Web search skills, favorite search engines, and domain expertise.
   c) The instructor explains the task execution procedure, hands printed guidelines to the user, including a short description of the application domain, and the tasks themselves. The user reads the tasks and asks the instructor for comments if necessary.
   d) The user studies standard on-site interface and query language descriptions of the designated search engine. The step can be skipped if he/she feels confident with the SE.
   e) The instructor explains ProThes interface to the user. The user performs simple tasks proving his/her ability to operate ProThes (the tasks include finding particular terms, navigating through thesaurus network, manual and semi-automatic query building, and executing queries).
2) Task execution stage. The user reports his/her readiness to perform the search tasks. The instructor runs logging utility and starts countdown. 15 minutes are allotted for each task. The user executes the task by copying & pasting URLs of relevant documents into a separate window. The user reports if he/she completes the task or gives up ahead of time.
3) Post-experiment stage
   a) The user fills out a final ProThes evaluation form.
   b) Informal discussion (optional).

User activities are logged with GURL Watcher utility (http://www.quicomm.com/gw_overview.html). Additionally, we register all ProThes-specific actions using a built-in logger.

Since we opted for a user-centered approach, we do not perform relevance judgment of the results reported by the user. We suppose that the user considers the listed URLs to be truly relevant from his/her point of view. It is crucial to convey the idea to the users that the study is aimed for *SE comparison*, not for examination on their Web search skills.

Users

Data were collected from 12 users who were researchers at the Institute of Engineering Science UD RAS or postgraduate students at Yekaterinburg universities. The main characteristics of the user group are shown in Table 2. All users except one indicated both Yandex and Google as their favorite SEs while one user only indicated Yandex. For all users the experiment was the first encounter with ProThes. All users are Russian native speakers.

Table 2. Users

Number of users	12
Mean age of users	27,58 years (range: 21–50)
Number of males/females	10/2
Number of students/graduated	4/8
Mean English language skills (self-estimation on a 5-point scale; 1=no knowledge, 5 = advanced)	3,42
Mean frequency of SE use (on a 5-point scale; 1=every day, 2=every 2-4 days)	1,17
Mean self-estimated Web search expertise (1= no knowledge, 5=expert)	3,42
Mean self-estimated domain expertise (1= no knowledge, 5=expert)	2,50

The experiments were conducted by two instructors (namely authors of the paper) on two sites on similar equipment with approximately equal Internet connection speed. Unfortunately we could not perform all experiments at similar times of day.

Results

Table 3 summarizes the main results obtained within the study.

Table 3. Basic research data

	G	Y	P
Mean queries per task	8,67	8,58	3,67
Mean query length in words	4,38	4,05	9,84
Null queries rate	0,20	0,11	0,21
Precise queries rate	0,37	0,19	0,13
Mean URL visited per nonempty query	1,46	- *)	2,14
Mean URL visited per result URL	6,37	- *)	3,54
Mean URL visited per task	10,08	- *)	6,57
Mean result URL for open-ended tasks	2,5	1,5	2,75
Mean time per task (min)	10,75	7,67	10,25
Uncompleted tasks rate	0,25	0,25	0,17

*) URLs visited in experiments 7-12 were not logged due to an unfortunate oversight.

Some comments on the mentioned parameters need to be added here. As 'null queries' we regarded queries, which delivered no results using either Google or Yandex, or both in case of ProThes. As 'precise queries' we regarded queries, which delivered nonempty lists with 50 or less results (sum of Google and Yandex responses in case of ProThes). 'Visited URLs' reflects the number of unique source pages viewed within an experimental task (the value does not include SE search forms and result pages).

It turned out, that search tasks' complexity was not equal.

The first task proved to be the easiest one. It took 5,13 minutes on average to complete the task. Both Yandex users completed the task within the first minute using a single query.

The 5th task turned out to be the hardest one. Only two users from 8 completed the task (remarkably both using ProThes, but the success can be explained through higher domain expertise as well).

In total 10 from 48 experimental tasks were not fulfilled. Moreover, 3 users gave up before reaching 15-minute limit (this is true for task #5 for all the three). This fact can be explained by either tiredness or low motivation of the users.

The hardest task (#5) produced also the highest query per task rates (means for Google and Yandex – 18,5 and 22,0 correspondingly; absolute maximum – 33 queries per task, user 12).

The fact-finding results consisted of a single URL (or none if the task was not accomplished). Open-ended tasks proved to be of different 'generality'. Thus, for the 6th task users reported twice as many results on average as in other open-ended tasks (#2, #4).

In general, fact-finding results showed more overlap in reported URLs. In case of both successful completions of the 5th task the same URL was indicated. In the 3rd task there were results from 3 different websites in 7 completed tasks. In the first task 5 URLs from 8 belonged to the same website. Results' overlap of the open-ended tasks is less significant.

As for specific ProThes features, users visited 2,08 thesaurus concepts per task on average. Queries built using ProThes included 79% thesaurus terms.

Table 4 summarizes post-search ProThes evaluation. It is to be noted that at least one user seemed to have a strong negative attitude towards ProThes – he rated all parameters with 1, which delivered an outlying result.

Table 4. ProThes' features evaluation (1=poor; 5=excellent)

Overall impression	3,00
Interface	2,25
Thesaurus visualization	2,83
Ease of use	3,17
Ease of learning	3,33
Performance	1,83
Query building	2,67
Thesaurus usefulness	2,92

Discussion

Analyzing research results we have to bear in mind that a small-scale user-centered study does not flatten individual characteristics of the users involved. So we should very carefully draw conclusions based on obtained quantitative data.

The user-centered approach cannot eliminate entirely users' subjectivity from the experiment. On the one hand, despite the given instructions several users still selected URLs without a thorough examination of the page content (maybe they behave the same way while solving real-life tasks). So, the quality (relevance) of the reported results may vary significantly from user to user. These concerns however do not affect interpretation of the data related to query-building phase. On the other hand, some users were low motivated or had even negative attitude towards experimentation.

Observations allow us to conclude that different users tend to employ different search strategies. Two distinct approaches are (1) to carefully work out a good query and (2) to start with a rough query and refine it gradually. ProThes is a highly specialized tool and obviously suits better for those users who employ the first approach.

During the experiments we noticed that most users had difficulties in switching between Russian, the language of the search tasks description, and English. It was crucial for many tasks since there are virtually zero documents in Russian on some task topics. The bilingual feature of ProThes can be helpful in these cases.

The uncommon ProThes interface has a higher 'acceptance threshold' than familiar interfaces of Google and Yandex. Moreover, the current implementation of ProThes is rather slow. These reasons can explain the significantly lower queries per task rates.

ProThes' thesaurus feature allows the users to build longer queries easier. Mean query length reflects this fact. However, long queries themselves are not an absolute good: many queries built using ProThes deliver null results or redundant response lists. These facts imply that query-loosening feature could be helpful in cases of too strict queries (this feature was proposed in [2] but has not been implemented yet). Note that average query length by Google and Yandex in the experiment is higher than commonly reported in comprehensive SE log studies (2-3 words). It can be explained by the particularity of the posed tasks.

'Partial matches' and use of Russian morphology can explain the lesser rate of null queries in case of Yandex.

The data on visited source pages are somewhat controversial. Moreover, due to an unfortunate oversight they are incomplete. On the one hand, ProThes users open external pages more rarely. It can be reasoned by longer Yandex' snippets presented in ProThes' results. A document description could suffice to make relevance judgment. Again, the figures can be explained by overall low performance of the system as well as by inconvenient way to access a source page from ProThes (as Java applet ProThes' client cannot start a web browser).

On the other hand, ProThes users visit more source pages with respect to each nonempty query.

The last lines of Table 3 show that despite of implementation shortcomings and 'novelty effect' ProThes can be quite useful for focused Web search.

It is to be noted that controlled manner of the user-centered approach decreases the utility of thesaurus – key feature of ProThes. The handed out task formulations already incorporated almost all necessary search terms. Thus, the user can skip the stage of verbalization of the information need, which appears challenging for many searchers where thesaurus becomes potentially very helpful.

Conclusion

Observations and user feedback gave us a good notion of future development of ProThes project.

First of all, overall system performance must be improved. Second, we must implement more standardized interface elements (e.g. operations with query tree analogously to Windows directory tree). Third, we should offer better thesaurus visualization.

As for developed user study scheme, we consider it to be a feasible and powerful technique for evaluation of search tools of different kinds. Controlled manner of the study allows us to fine-tune the method easily, target different aspects of search process, as well as find balance between user subjectivity and large amount of data to be collected and processed.

Acknowledgments

The research was supported in part by the Russian Fund of Basic Research, grant # 03-07-90342.

We would like to thank all volunteers for their participation in the study as well as for valuable comments on ProThes system.

Appendix. Search tasks (originally in Russian)

1) How many accuracy classes for printed circuit boards are defined?
2) Find documents on algorithms for edge detection applied to printed circuit board images in automatic optical inspection tasks.
3) Minimal line width for *Orion* printed circuit board automatic optical inspection system.
4) Find documents on vectorization algorithms for printed circuit board images (image vectorization means transformation of bitmap image into a vector image).
5) Maximal printed circuit board size that can be tested using *Discovery 6* automatic optical inspection system.
6) Find descriptions of automatic optical inspection systems which use X-ray sources.

References

1. Aula A. Query Formulation in Web Information Search // Isaías P. & Karmakar N. (Eds.) Proceedings of IADIS International Conference WWW/Internet 2003, volume I, 2003. P. 403-410.
2. Braslavski P., Shishkin A., Alshanski G. 3 in 1: Meta-Search, Thesaurus, and GUI for Focused Web Information Retrieval // Digital Libraries: Advanced Methods and Technologies, Digital Collections. Proceedings of the 6th National Russian Research Conference, September 29 - October 1, 2004, Pushchino. P. 135-140.
3. Buzikashvili N. IR interactions: Myths, Reality and Basic Principles. [Poiskovoe vzaimodeistvie: mify, real'nost' i bazovye prinzipy] (in Russian) // Digital Libraries: Advanced Methods and Technologies, Digital Collections. Proceedings of the 6th National Russian Research Conference, September 29 - October 1, 2004, Pushchino. P. 226-235.
4. Korobkina N. The Research of Users' Behaviour in the Tasks of Scientific Search [Issledovanie povedeniya grupp pol'zovatelei v zadachakh nauchnogo poiska] (in Russian) // Digital Libraries: Advanced Methods and Technologies, Digital Collections. Proceedings of the 6th National Russian Research Conference, September 29 - October 1, 2004, Pushchino. P. 290-297.
5. Rose D. E., Levinson D. Understanding User Goals in Web Search // Proceedings of the Intl. World-Wide Web Conference 2004, May 17-22, 2004, New York, NY USA. P. 13-19.
6. Russian Information Retrieval Evaluation Seminar, http://romip.narod.ru
7. Spink A. A user-centered approach to evaluating human interaction with Web search engines: an exploratory study // Information Processing and Management 38 (2002). P. 401-426.
8. Tombros A., Sanderson M. Advantages of Query Biased Summaries in Information Retrieval // Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia. P. 2-10.
9. White R. W., Jose J.M., Ruthven I. A task-oriented study on the influencing effects of query-biased summarisation in web searching // Information Processing and Management 39 (2003). P. 707-733.